Institut für Experimentelle Mathematik

Conference on

Retrodigitization of mathematical journals and their integration into searchable digital libraries

01.08. – 03.08.2000

Organizers: R. Fateman (Berkeley)
G. Michler (Essen)
E. Mittler (Göttingen)

Prof. Dr. G. Michler, Ellernstraße 29, 45326 Essen
Tel. (+49) (0) 201 183-7639 , Fax (+49) (0) 201 183-7625

Do you have ssh on X-stations.

Conference on

Retrodigitization of mathematical journals and their integration into searchable digital libraries

Institute for Experimental Mathematics, Essen

August 1-3, 2000

Revised

Program

Tuesday, August 1

9.15	-	9.20	**Opening**	
9.20	-	10.05	**R.K. Dennis** Ithaca	*A global mathematics library*
10.10	-	11.00	**O. Ninnemann** Karlsruhe	*Jahrbuch Project and the Electronic archive of mathematical literature*
			Coffee-Break	
11.30	-	12.20	**S. Thomas** Ann Arbor	*JSTOR's experience with retrodigitization of mathematics literature*
12.30	-	13.00	**M. Höppner** Bielefeld	*What is the digital library NRW good for a mathematician?*
			Lunch-Break	
14.30	-	15.20	**M. Köttstorfer** Linz	*DIEPER – Digitised European Periodicals*
15.30	-	16.20	**R. Fateman** Berkeley	*A multivalent document system and its use in the Berkeley library project*
			Coffee-Break	
16.50	-	17.40	**M. Fock-Althaus** Berlin	*AGORA the new digital library system of SRZ and the SB in Göttingen based on XML/RDF format*
17.45	-	18.30	**M. Enders** Göttingen	*The Digitization process for mathematical journals – An introduction to the workflow at the GDZ*

Wednesday, August 2

9.15	-	10.00	**D. Nastoll** Essen	*MILESS*
10.05	-	10.45	**E. Mittler** Göttingen	*Retrodigitized and born digital material – how to bring it together?*
			Coffee-Break	
11.15	-	12.05	**M. Okamoto** Nagano	*Recognition of mathematical symbols*

ad hoc

ad - bc

12.10	-	13.00	**K. Yokota** Tokio	*Layout analysis of mathematical documents*
			Lunch-Break	
14.30	-	15.00	**G. Michler** Essen	*Retrodigitization of "Archiv der Mathematik"*
15.00	-	15.45	**R. Staszewski** Essen	*Analysis of mathematical documents*
			Coffee-Break	
16.15	-	17.00	**R. Nörenberg** Essen	*Recognition of mathematical formulas*
17.10	-	18.00	**B. Luchner** Basel	*Das Cross-Ref-Projekt. Eine Initiative führender Wissenschaftsverlage auf der Basis des DOI-Systems*

Thursday, August 3

9.15	-	10.00	**M. Suzuki** Fukuoka	*A design of optical recognition system of mathematical documents – An experimental implementation*
10.00	-	10.30	**Y. Etou** Fukuoka	*Optical recognition of mathematical expressions using spanning tree of minimal cost*
			Coffee-Break	
11.00	-	11.45	**F. Tamari** Fukuoka	*On the segmentation of Japanese Area/Mathematical Area*
11.45	-	12.00	**M. Nakayama** Fukuoka	*Demonstration*
12.00	-	12.45	**R. Fateman** Berkeley	*Experiments in creating and using scanned journals on-demand*
			Lunch Break	
14.30	-	15.20	**G.Schneider** Göttingen	*Aspects of long term preservation of digital libraries*
			Coffee-Break	
15.50	-	17.00	**Chair: Dr. J. Bunzel** DFG Bonn	*Final discussion*

Organizers: Prof. R. Fateman, Prof. Dr. G. Michler, Prof. Dr. E. Mittler
The conference is supported by the Deutsche Forschungsgemeinschaft

Teilnehmerliste
Tagung "Retrodigitization of mathematical journals and their integration into searchable digital libraries" vom 01. - 03.08.2000

Organisatoren:
Prof. R. Fateman (Berkeley)
Prof. Dr. G. Michler (Essen)
Prof. Dr. E. Mittler (Göttingen)

Dr. H. J. Becker
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen
e-mail: becker@mail.sub.uni-goettingen.de

M. Bieber
Anwendungszentrum Neuro-Informatik e.V.
Hagenower Str. 73
19061 Schwerin
e-mail: bieber@ani.de

A. Bilo
Bibliotheksdirektor
Universität GH Essen
45117 Essen
e-mail: bilo@bibl.uni-essen.de

Dr. J. Bunzel
Wiss. Bibliothekswesen
Deutsche Forschungsgemeinschaft
Kennedyallee 40
53175 Bonn
e-mail: Juergen.Bunzel@dfg.de

Prof. R. K. Dennis
Department of Mathematics
Cornell University
524 Malot Hall
Ithaca, New York 14853-7901
USA
e-mail: dennis@rkd.math.cornell.edu

M. Enders
Niedersächsische Staats- und
Universitätsbibliothek Göttingen - GDZ
Papendiek 14 (Altbau)
37073 Göttingen
e-mail: enders@mail.sub.uni-goettingen.de

Y. Etou
Graduate School of Mathematics
Kyushu University 36
6-10-1, Hakozaki, Higashi-ku
Fukuoka 812-8581
Japan
e-mail: etou@moon.ap.kyushu-u.ac.jp

Prof. R. Fateman
Computer Science Division
721 Soda Hall
University of California, Berkeley
Berkeley, CA 94720
USA
e-mail: fateman@cs.berkeley.edu

Dipl.- Kfm. M. Fock-Althaus
Satz-Rechen-Zentrum
Bessemerstr. 83-91
12103 Berlin
e-mail: m.fock-althaus@srz-berlin.de

Dr. S. Göbel
Fachinformationszentrum Karlsruhe
Abteilung Mathematik und Informatik
Franklinstr. 11
10587 Berlin
e-mail: goebel@zblmath.fiz-karlsruhe.de

Dr. A.-K. Hillebrand
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen
Am Faßberg
37077 Göttingen
e-mail: ahilleb2@gwdg.de

Dr. M. Höppner
Bibliothek der Universität Bielefeld
Postfach 10 02 91
33502 Bielefeld
e-mail: hoeppner@ub.uni-bielefeld.de

Dr. M. Karbe
Verlag Walter de Gruyter & Co.
Postfach 30 34 21
10728 Berlin
e-mail: m.karbe@degruyter.de

M. Köttstorfer
Computer Science for the Blind
University of Linz
Department of Computer Science
Altenbergerstr. 69
4040 Linz
Österreich
e-mail: marco@aib.uni-linz.ac.at

B. Kratzer
Institut für Experimentelle Mathematik
Universität GH Essen
Ellernstr. 29
45326 Essen
e-mail: gradkol@exp-math.uni-essen.de

Dr. B. Lix
Hochschulrechenzentrum
Universität GH Essen
Schützenbahn 70
45117 Essen
e-mail: lix@hrz.uni-essen.de

Bernd Luchner
Birkhäuser Verlag
Viaduktstraße 40-44
CH-4051 Basel
Schweiz
e-mail: luchner@birkhauser.ch

F. Lützenkirchen
Hochschulrechenzentrum
Universität GH Essen
Schützenbahn 70
45117 Essen
e-mail: luetzenkirchen@hrz.uni-essen.de

Prof. Dr. G. Michler
Institut für Experimentelle Mathematik
Universität GH Essen
Ellernstr. 29
45326 Essen
e-mail: archiv@exp-math.uni-essen.de

D. Mirrouch
Institut für Experimentelle Mathematik
Universität GH Essen
Ellernstr. 29
45326 Essen
e-mail: mirrouch@exp-math.uni-essen.de

Prof. Dr. E. Mittler
Direktor
Niedersächsische Staats- und Universitätsbibliothek Göttingen
37070 Göttingen
e-mail: mittler@mail.sub.uni-goettingen.de

Dipl.-Inf. C. Mönch
Fachbereich Informatik
Universität Frankfurt
Postfach 11 19 32
60054 Frankfurt
e-mail: moench@uni-frankfurt.de
cristian@tm.informatik.uni-frankfurt.de

M. Nakayama
Fukuoka University of Education
Department of Mathematics
Akama 729, Munakata
Fukuoka 811-4192
Japan
e-mail: nakayama@crown.fukuoka-edu.ac.jp

D. Nastoll
Hochschulrechenzentrum
Universität GH Essen
Schützenbahn 70
45117 Essen
e-mail: nastoll@hrz.uni-essen.de

O. Ninnemann
Abt. Berlin FIZ Karlsruhe
Zentralblatt MATH
Franklinstr. 11
10587 Berlin
e-mail: olaf@zblmath.fiz-karlsruhe.de

Dr. R. Nörenberg
Institut für Experimentelle Mathematik
Universität GH Essen
Ellernstr. 29
45326 Essen
e-mail: noerenbg@exp-math.uni-essen.de

Masayuki Okamoto
Department of Information Engineering
Faculty of Engineering
Shinshu University
500 Wakasato, Nagano City 380
Japan
e-mail: okamoto@wwwserve.cs.shinshu-u.ac.jp

Prof. T.R. Reed
Department of Electrical and Computer Engineering
University of California
Davis, CA 95616
USA
e-mail: reed@ece.ucdavis.edu

Prof. Dr. G. Schneider
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen
Am Faßberg
37077 Göttingen
e-mail: gerhard.schneider@gwdg.de

Dr. R. Staszewski
Institut für Experimentelle Mathematik
Universität GH Essen
Ellernstr. 29
45326 Essen
e-mail: reiner@exp-math.uni-essen.de

Dr. A. Stein
Institut für Algebra und Geometrie
Universität Halle-Wittenberg
Theodor-Lieser-Str. 5
06120 Halle
e-mail: stein@coxeter.mathematik.uni-halle.de

Prof. Masakazu Suzuki
Graduate School of Mathematics
Kyushu University 36
6-10-1, Hakozaki, Higashi-ku
Fukuoka 812-8581
Japan
e-mail: suzuki@math.kyushu-u.ac.jp

Prof. Fumikazu Tamari
Fukuoka University of Education
Department of Mathematics
Akama 729, Munakata
Fukuoka 811-4192
Japan
e-mail: tamari@fukuoka-edu.ac.jp

S. Thomas
JSTOR
301 E. Liberty St. Suite 310
Ann Arbor, MI 48104
USA
e-mail: spencer@umich.edu

Dr. B. Tröger
Universitätsbibliothek Essen
Universität GH Essen
45117 Essen
e-mail: troeger@bibl.uni-essen.de

Dr. K. Waki
Institut für Experimentelle Mathematik
Universität GH Essen
Ellernstr. 29
45326 Essen
e-mail: slwaki@exp-math.uni-essen.de

Dr. K. Yokota
Computer & Multimedia Section 4
Computer & Multimedia Hardware Design Department
Ome Operations
Toshiba Corporation Digital Media Network Company
Suehiro-cho 2-9, Oume-shi
Tokio 198-0025
Japan
e-mail: kazuaki.yokota@toshiba.co.jp

Dr. W. Ziegler
Universität Jena
Phys. Astron. Fakultät
Max-Wien Platz 1
07743 Jena
e-mail: Wolfgang.Ziegler@rz.uni-jena.de

Abstract:

The Digitization Process for Mathematical Journals

An introduction to the workflow at the GDZ

Markus Enders, GDZ / SUB

The GDZ

The "Göttinger Digitization Center" (GDZ) was founded in 1997 with the task to support the DFG-program "retrospective digitization of library materials" and to coordinate national efforts concerning digitization. Therefore the GDZ is involved in several projects:

- Digitization of Travel Literature and North-Americana, books from the 18th and 19th century
- Digitization of the "Göttingische Gelehrte Anzeigen"
- Digitized European PERiodicals (DIEPER), which is a project funded by the European Commision to establish a European wide database of digitized periodicals as a common access point to the digitzied journals of all project partners. This will not only include searchable bibliographic metadata of journals or articles, but also searchable fulltext of selected volumes.
- Digitization of the Gutenberg Bible (1999)

To provide one access point to all the digitized material and to build a solid basis for other, future projects, it was decided to develop a Document Management System (DMS) together with the company Satz-Rechen-Zentrum in Berlin.
Right now, this Document Management System provides access to over 1200 digitized works (monographs, volumes....) with almost half a million images.

Workflow

Currently we established a workflow at the GDZ which includes a broad range of activities: From scanning the material, capturing the metadata, till providing the material on our Document-Server and storing it for long time preservation (archieving). Our workflow in detail:

- Selection of the material by librarians
- Capturing of the structure (e.g. of a book): This is done in Microsoft Excel. We use a special spreadsheet, which contains all information about chapters, articles etc. of a journal or a monograph.
- Scanning of the material: We use a special software called SRZ-ProScan, which allows us to have a great throughput: E.g. two pages can be scanned with just one scan, pages are numbered automatically, Bibliographic Information is inserted in the Header of each TIFF-Image automatically. Currently we are scanning with 600dpi, 1bit for normal material. Images or maps are scanned with 4 or 8 bit.
- Quality control of the images: The quality of the images are checked (to dark, to light, something missing) Usually we only use the Excel spreadsheet, to check, if pages are missing / were not scanned. Only in some cases, the original book is needed.
- If the images pass our quality control, the book is marked as "already digitized" and returned to its original location. If the images do not pass the control, they are returned to the scanning operator, with a report of the missing/bad images.
- Archieving the scanned images on CD-ROM.
- Images are enchanced using speckle and despeckle algorithms. The program we use right now (PixEdit) allows batch proccessing.
- Capturing of metadata: Capture titles and authors of chapters and articles as written in the Table of Contents. Currently we use three different methods to capture the metadata:
  - Giving the images (of the Table of Contents) away to other companies, who capture the metadata and insert them into the Excel-spreadsheet
  - Capturing the metadata ourselves with support of OCR (using Finereader).

- We had some experiments to extract the metadata of each chapter/article (title, author, page) directly from the OCR output and insert it into the Excel spreadsheet automatically. This was quite successful, but only seems to be useful for journals, because every variation in the Table of contents needs a modification of the program.
- Convert the Excel spreadsheet to XML: At the GDZ we developed a program called DigDoc, which converts the information from the Excel spreadsheet to a XML-file.
- Capture Metadata from our online catalog: Most of the material we digitize is already available in our library online catalog (PICA). So we could use the metadata on journal or monograph level and include it into the XML file, we produced with the DigDoc program. This is done by a java program we developed at the GDZ, called Pica2XML.
- As the result of our workflow, we get an XML file, which contains all the metadata and the structure of the digitized work. This file is imported into the Document Management System called AGORA.
- The final work (images and the xml-file) is archieved on CD-ROM

Throughout the whole process a documentation is needed, so that each employee knows, in which state the material is. Especially when working with a larger team on digitization this becomes very important, because not a single person in the team has an overview over the whole production of digitized material. A workflow system have to give at least the following information:

- Which is the material a team member can work with and where to find this material on our storage. Especially when you have several digitization projects it is not easy, to find the images or the Excel spreadsheet immediately.
- Statistics about a certain project is needed (how many works are already online, where are the other works in the workflow....)
- Where is a certain work archived (on which CD-ROM)
- The workflow system must have information on the responsibility of a team member for certain material or certain steps. If freelance workers are employed (e.g. for scanning), the system may deliver information on the payment they receive.
- The system must collect all the information and give easy access to all files (images, files containing metadata, files containing fulltext...) belonging to a work. Searching the appropriate files on the storage by hand (in the file system) takes too long for extensive production.

Currently we have not implemented a workflow system which covers the whole workflow. Only the first steps from scanning to archieving the Images are managed with a database (in MS-Access). This system is also used to determine the payments for our freelance workers (scanning operators).
All further steps (from capturing the metadata to archivieing the images and metadata together) are only recorded in Excel spreadsheets. Due to the large amount of data, we have to use several spreadsheets for storing all information.

The creation of fulltext is not yet implemented anywhere in our workflow. It is still a field of experimentation and research. The mathematical journal we digitized for the DIEPER project and will not be OCRed, but their fulltext will be captured by an external company. Due to several problems (bad print, bad scans and mathematical formulas), we are just planning to use the fulltext for searching purposes, not for display. That's why we just capture the position of the formulas in the text, but not their contents.

Data formats

At the GDZ a file format was developed for storing metadata and structure of a digitized work in one file. This format is based on RDF and XML, two common known standards. The reasons for using XML and RDF were:

- Interoperability, with XSLT style sheets and appropriate XSLT-processors xml files can be converted to other xml based file formats.
- common known standard, meanwhile a large amount of programming libraries are available
- support of several character encodings, support of Unicode (utf-8 and utf-16 encoding)

With the involvement in the DIEPER project and the demand to support fulltext, this file format has to be updated and extended. Together with other DIEPER partners the GDZ developed a slightly modified file format which integrated Dublin Core to the metadata section. For compatiblity reasons, it was decided not to include the fulltext into the same file, but to use another common known standard: TEI (Text Encoding Initiative) in its XML version. The fulltext tagged according to the TEI standard is stored in a separate file. Both files (the RDF/XML containing the metadata and the TEI/XML containing the fulltext) are linked via unique identifiers.
Currently this new file format is not implemented at the GDZ and cannot be produced directly. But the use of XML technology as style sheets and XSLT processors, allows us to convert our older RDF/XML files to the newer version.

Y. Etou (with M. Suzuki)

Optical Recognition of Mathematical Expressions Using Spanning Tree of Minimum Cost

In this lecture, I will describe our new method to recognize mathematical expressions. The problem of adaptation of OCR systems for scientific documents including mathematical expressions is, roughly speaking, devided into the following three problems:

1. Adaptation of character recognition engine for mathematical symbols,
2. Segmentation of the text parts and mathematical expression parts,
3. Recognition of the structure of mathematical expression.

This lecture concenrns the third problem. There are many works on the structure analysis of mathematical expressions. Some of them uses less the recognition results of characters and symbols, the others assume the correct recognition results. Since it is very hard to find a practically usable method of structure analysis without using recognition results of characters and symbols, we took the latter way. In our laboratory, we once implemented an analyzer of mathematical expressions assuming the recognition results of characters and symbols [1]. However, the recognition results of OCR engine always containes errors in some rates, which leeds sometimes to fatal errors in the structure analysis of expressions. There are also some pairs of symbols with capital letter and small letter with same or similar shape such as "S" ans "s", "C" and "c" etc., which causes also errors of the structure analysys, since the structure of mathematical expressions are recognized, in many cases, by using the relative sizes and the positions of the characters and symbols. Another problem for the structure analysis is the differences of the printing styles between various mthematical documents: variation of the positions and the sizes of symbols and subscripts etc.

In order to overcome these difficulty, we introduced a new algorithm of the structure analysis which improves the robustness against

- the variation of the printing styles of documents and
- the errors of the recognition of the characters.

The system first constructs a virtual link network with vertices representing the the candidates of the recognition results of the characters and symbols, linked each other by edges with

- the label representing the possible relationship of the pair of characters or symbols: subscript, superscript or on a same line, etc., and
- the cost of the relation.

Then, the system outputs the spanning tree of the network with the minimum cost as the result of the structure analysis of the mathematical expression. The advantage of this method is that small local errors of the recognition are recovered automatically by the total cost of the recognition tree.

The costs of the links are calculated using the density functions of the normal distributions determined by statistical data of the relative sizes and the positions

obtained from our database for each categorical pair of characters or symbols, in order that the algorithm cope with the variation of the printing styles of the documents.

In the lecture, we report our experimental results for mathematical expressions extracted from various mathematical journals.

[1] K.Inoue, R.Miyazaki, M.Suzuki, "Optical Recognition of Printed Mathematical Documents", Proceedings of the Third Asian Technology Conference in Mathematics, Springer-Verlag, 280–289, 1998

R. Fateman

Experiments in creating and using scanned journals on-demand

We will describe on-going experiments at the University of California at Berkeley in which users of the library will be encouraged to make digital scans of articles of interest using a hand-held HP Capshare scanner. These can then be send as multi-page TIFF files to a web-based server which cleans up the image, runs OCR and indexing programs, and inserts the resulting document in a private or public database. We will also explore the usefulness of having laptop computers in the library serving as an intermediate storage medium for scanned pages. The Capshare scanner itself can store only about 40-50 pages before it must be emptied. The software running in the web server is a mixture of commercial and locally-written programs.

M. Fock-Althaus

Agora the new digital library system of SRZ and the SUB in Göttingen based on XML/RDF format

Current publications are more and more available in electronic form. Nevertheless there is still a clear predominance of printed material within libraries and other institutions. In order to faciliate online access to these informations digitization is necessary. The german research foundation (DFG) started in 1997 a programm of retrodigitization. A technical work group was set up and defined an extended format based on tiff-images for the exchange of data between the several liberian institutions. SRZ Berlin is a commercial service company with some background in the field of electronic publishing and document management. In 1998 we took the decision to participate to the programm of the DFG. Based on an intensive working relation between the liberians of the SUB in Gttingen and our software engineers we looked at and checked the commercial DMS products of several vendors to the needs of a library system for presenting digitized and electronic documents. The system had to be open, plattform independend and integrating standards like JAVA, XML and ODBC. After our investigation we decided to create our own solution, the Agora System. The DMS Agora is now available in german in version 1.0. An English version will be ready in September 2000.

With the DMS Agora we have also developed components, like SRZ ProScan Book, Agora XML-Editor and Agora Konverter for digitizing, indexing, converting and structuring of paper and microforms. SRZ can now present a complete workflow from printed material to electronic presentation of the contents in the web.

The Agora system contains in principel of four depending moduls:

- Agora Admin (central modul for all operations)
- Agora Database
- Agora Servlet
- Agora and Verity search engine

The lecture will cover the demonstration and functionality of the Agora system based on some demo documents. Further to this there will be an explanation of the usage of the different formats RDF/XML and TIE/XML through Agora.

A.K. Hillebrand

Active documents. Using XML with the new release of IBM DB2 Digital Library

The lecture will give a short introduction to MELISSA, the joint project of GWDG and Bayerische Staatsbibliothek, which will be accomplished in cooperation with the MILESS project in Essen. MELISSA is the name for a distributed multimedia system for publishing and editorial work. The aim of this project is to build a digital library for the historical sciences which can be used for web-based retrieval and usage of multimedia documents as well as for web-based creating, editing and reviewing of documents. The objective is to establish a virtual faculty for people doing research in the historical field and to support research and education at real university departments. MELISSA consists of two main components: a multimedia encyclopedia for the historical sciences and a system for managing university multimedia publications. Of course, the whole system is designed to be applicable for other faculties as well. The projected solution involves as a software basis IBM DB2 Digital Library (DL) and the MILESS application developed for DL. One major task of the project is the adaption of MILESS' document model to complex scientific document structures. This is where XML comes into operation: Taking into account internatonal standards like Dublin Core and RDF, structures of existing document types like monographs, essays and encyclopedic articles as well as audio and video documents etc. will be mapped to DL using XML. In this regard, the ability to build combined documents is of great importance. A combined document may, for example, consist of a thesis and a 3D visualization to illustrate the thesis. Users may want to find both objects merged to one document when they search for theses on a given subject. They also may want to get the 3D visualization as a document on its own when they search for all 3D visualizations contained in the digital library. Moreover, a special DTD will be supplied for the highly structured encyclopedic articles. This DTD - together with a web-based editor - will enable authors to import their articles as XML files without having to acquire special skills. The articles will be automatically prepared for usage with a webclient including an intelligent linking of articles and related materials and an automatic highlighting of links directly in the article text. The lecture focusses on the modeling of documents and their representation in the MELISSA digital library. Furthermore, it provides information about how DL supports XML on the level of the text search engine as well as on the level of the object oriented application programming interfaces (OO APIs), and what effort has to be spent on the development of additional modules which enable mapping of complex document structures using XML and support import of XML data.

M. Höppner

What is the Digital Library North Rhine-Westphalia good for a mathematician?

Building up the Digital Library North Rhine-Westphalia as a portal to the world of science is a project of the appropriate state authority, the Ministerium fr Schule und Weiterbildung, Wissenschaft und Forschung des Landes Nordrhein-Westfalen. Started in April 1998 the project is carried out in co-operation between the Hochschulbibliothekszentrum (HBZ) and the universities and polytechnics in North Rhine-Westphalia. Until the end of 1999 Bielefeld University Library was responsible for the whole project, now administration and daily routine have been passed to the HBZ while all further developments remain in Bielefeld.

The aim of the Digital Library NRW is to offer a seamless integration of search for, record of and access to information under one common user interface. This means in particular:

- uniform access to heterogeneous and dispersed resources under one user interface, including even databases based on different software and on different technical platforms,
- simultaneous bibliographic recording of literature in electronic and in printed form,
- complete recording of library holdings for printed documents and direct access to all (licensed) electronic documents,
- immediate transfer of records to document delivery services,

thus ensuring a comfortable access to scholarly information for students and scientists.

The concept resulting of these objects is based on several cornerstones:

- Record of information: The record of literature for printed publications in library holdings as well as for electronic forms free available in the Internet or purchased from publishing houses is combined electronically by using web based search engines, databases with indexing data and retrieval in heterogeneous databases under on common user interface.
- Access to information: The user is offered different forms of access for every single record given in the integrated search. He will get the information whether he can obtain the recorded document in printed form in his local library or online in electronic form accessible directly from the record or via a document delivery system.
- Gathering of information: Scientific publications free available in the Internet are recorded in an own catalogue. The bibliographic data of these sources is partially registered manually on intellectual bases and is partially collected (semi-) automated.

To realize these conceptual cornerstones the Digital Library NRW is build upon several technical components: a central access system for using the electronic services of the Digital Library under one common user interface, a central access control and monitoring system, a central billing and accounting systems (although other local accounting systems can also be used) and document delivery systems e.g. for journals articles (an electronic system for interlibrary loan of monographs is just under development and will be tested within this year).

Accessing the Digital Library NRW every user will get a so-called local view on all services that are licensed to the campus of the university or polytechnic he

belongs to. Local sights are already implemented for twelve institutions, others will follow. An external user who belongs to no academic institution in North Rhine-Westphalia will get a special view on all services free of licence or of costs. Recently the discussion about the possibilities how to combine the Digital Library NRW and the local Websites of the universities and polytechnics has raised more and more attention. As the first institution the Bielefeld University Library has integrated the Digital Library NRW in her own local Websites by creating smooth transitions between both systems thus offering her users an easier and transparent access to all local and central electronic services.

As the Digital Library is a dynamic system for information supply of the universities and polytechnics it certainly requires a continuous development in technology and extension of content. For example user profiles will be implemented in the near future. Within the integrated search a user interested in a special subject like mathematics will have the possibility to choose the databases relevant to his research and store the term performed as an own search profile. He will be informed periodically about new documents referring to this profile.

Besides the technical development of the Digital Library NRW the ministry also funded the acquisition of commercial electronic publications for all subjects, especially full text databases or training material with interactive and multimedia elements. Most of the funding was spent for the subscriptions of all the electronic journals by the publishing houses of Elsevier and Springer which will be accomplished by the electronic journals of Academic Press and Kluwer soon. Of course, the journals of these publishers are highly interesting for every mathematician. The following databases included in the Digital Library NRW are also of fundamental interest for a mathematician:

- MATH Database (Zentralblatt für Mathematik)
- MATHDI Database (Zentralblatt für Didaktik der Mathematik)
- MathSciNet (Mathematical Reviews)
- Sigma (Math Net)
- INSPEC including Computer and Control Abstracts (IEEE)
- Science Citation Index including Current Contents (Institute of Scientific Information)

Among the further publications offered in the Digital Library NRW at least the following ones are relevant for a mathematician:

- DeskTop Bronstein: Taschenbuch der Mathematik (Verlag Harri Deutsch)
- DeskTop Stöcker: Taschenbuch mathematischer Formeln und moderner Verfahren (Verlag Harri Deutsch)
- CINDERELLA, the Interactive Geometry Software (Springer-Verlag)
- Standard Math Interactive (Springer-Verlag)
- REVBOOK: Ein wiederbelebbares Buch zur Statistik (Universität Bielefeld, Fakultät für Wirtschaftswissenschaften)

The content of the Digital Library NRW will be extended every year and all of you are invited to suggest further relevant material for acquisition or even to create new material as a follow up from your research or teaching and present it to the public by the Digital Library NRW.

The core system of the Digital Library NRW may be accessed by http://www.digibib-nrw.de/ (use "login" followed by the choice "Gastzugang"), for a complete integration of the Digital Library NRW into the local service of a library see http://www.ub.uni-bielefeld.de/

M. Köttstorfer

DIEPER Digitised European Periodicals

Partners from ten European countries have joined to build a virtual network and a central access point to make accessible periodicals that have been retrospectively digitised in Europe or anywhere else in the world.

Some questions DIEPER will try to answer:

- Which printed periodicals have been retrospectively digitised in Europe and anywhere else in the world?
- How may I access digitised periodicals without searching a number of different internet sites?
- Where may I browse selected periodicals and to what extent are texts searchable?
- How can digitised periodicals be preserved and kept accessible in permanence?
- How will rights of authors, publishers and libraries be protected?
- How will libraries subscribe to digitised periodicals?
- How will online licensing and accounting be done?

Technical progress has made digitising of printed documents an option for most modern libraries. Digitising is going on at a rapid pace already. In most cases the selection of documents, the technical standards applied, and the options for access by the user follow the individual decision of the library. In spite of the costs involved this results in considerable duplication of effort (even the same items digitised at different places) and above all, no clear strategy can be offered to the user, which may help him to find out, whether the document he needs is available as digital full text. This is a particular disadvantage regarding back issues of important research periodicals, where users are looking for fast access to relatively short documents (articles).

DIEPER addresses the need for a central access point where all digitised periodicals shall be recorded. This shall be devised as a register built on the model of the European Register of Microform Masters. Records of the register will be linked to reliable and comprehensive archives of periodical literature working at different sites throughout Europe. Those archives will of course be equally accessible from other points (e.g. WWW pages and bibliographic databases) but the register will be the only place, where it is attempted to file all the relevant records. In addition a search machine accessible from the register will allow to do a fulltext search of the articles in digitised periodicals or at least a fulltext search of the tables of contents.

The DIEPER partners will contribute to build a virtual library of periodicals by scanning a few journals according to de facto standards and to demonstrate the technichal feasability of optimum online access and retrievability. At the same time they will make accessible any already existing digitised periodical by including its record in the register and by making the documents' indices available to the search machine. In establishing this infrastructure it is hoped to develop a uniform strategy for retrodigitising throughout Europe, that is based on a minimum standard for access and retrieval.

While dramatically improving access to older periodical literature, offering it to users in digitised form is a direct contribution to protecting the original items and preserving the printed information in another format. Availability of periodicals through DIEPER will increase use - continuous use will make the online hosts aware of the need to keep their digitised periodicals accessible.

DIEPER helps to fill the gaps in existing library collections and may in the long-term contribute to reduce operating costs of libraries. Just as important as this is its role in providing a point of global information for those who want to contribute to build the virtual electronic library without risk of duplicating effort.

B. Luchner

Das Cross-Ref-Projekt. Eine Initiative führender Wissenschaftsverlage auf der Basis des DOI-Systems

CrossRef is the name of a reference linking initiative started by a group of leading STM publishers in December 1999. The initiative aims to provide a service that will link reference citations to the online content that those references cite. The CrossRef service is currently the leading application based upon the new persistant Digital Object Identifier (DOI). The CrossRef service links the online publications on different servers from different publishers and could evolve to an indispensable mechanism working in the background of any digital library. My lecture will give an overview on the concepts of CrossRef and DOI and on the online ressources that already were implemented. I also will report on the state of the art of CrossRef and DOI at the online service of the Springer Group of companies LINK.

G. Michler

A prototype of a combined digital and retrodigitized searchable mathematical journal

In this lecture a survey is given on the programs used for the retrodigitization of 6 volumes of the mathematical journal "Archiv der Mathematik" appeared in the period 1993 to 1995. The result may be considered to be a prototype for a mathematical text recognition system. Furthermore, methods have been developed to incorporate the retrodigitized texts into a digital library database. Thus it is possible to link the retrodigitized back issues with the recent digital issues of a mathematical journal. Furthermore, full searchability within the retrodigititzed texts has been achieved.

In this project also the recognition problem of mathematical formulas has been addressed. Using the programs of Professor Okamoto's study group (Nagano) most of the mathematical formulas contained in the 6 volumes of the Archiv der Mathematik have been recognized and transferred into latex form.

The linkage problem has been solved by means of the new programs for the recognition of the layout of the first page of a scanned article. It produces an XML-file of the bibliographic data of such an article. These bibliographic meta-data allow the integration of the MVD format into the digital library database MILESS. The multivalent document format MVD was developed by T. Phelps (Berkeley); it is another main ingredient of the retrodigitization program system.

Another program has been developed for recognizing the references of a scanned article. It provides the XML-files of the articles listed in the bibliography of such an article. These articles contained in the digital library can be read on screen together with the first article.

Most of these programs have been developed by Dr. R. Staszewski (Essen) and Dr. R. Nörenberg (Essen). They will demonstrate these programs and give further information in their lectures at the meeting.

M. Nakayama

Demonstration of the segmentation of Japanese area/Mathematical area and the structual analysis of mathematical expressions

We are developing an OCR system which is able to recognize mathematical symbols and analyze symbol arrangement with Suzuki in Kyushu University, Fukuda in Oita University and Toshiba Co.,Ltd. in Japan.

In our demonstration, we propose the segmentation of Japanese area/Mathematical area in Japanese text books including mathematical formulae. Tamari in Fukuoka Univ. will talk about the segmentation of Japanese area/Mathematical. So here we will explain about the system which analyzes mathematical expressions whose arrangements are two dimensional. We use a top down method and bottom up method to analyze scanned image data which have mathematical symbols.

We divides mathematical expressions into some blocks such as index area, fraction area and root area and analyzes the structure. By processing those equations individually, we can analyze the structure easily. An in-line or a displaystyle mathematical expression is divided into multiple blocks using features such as large fraction line, Σ, $\int$ and parenthesis. The connection relations of these blocks are called same levels. For the characters such as the alphabet, the division is judged by using the normalized circumscription rectangle, since there is a difference of their size among a,b,g, and so on.

Furthermore, for each divided block, above mentioned division is carried out. It is repeatedly done, until the character ion the base line in the block becomes just one. The except ones for characters on the base line are considered the characters of the suffix region. By this method, the error of the conformational recognition that suffixes, etc. gives is settled in one block.

D. Nastoll, F. Lützenkirchen

MILESS - The MultiMedia Digital Library Essen

MILESS (MultiMedia Learning and Teaching Server Essen) was started as a joint project between the Computer Center and the Central Library of Essen University, together with the two pilot departments of linguistics and physics in 1998. Today its a joint service of the three central units, the IT-Center, the Central Library and the Media Center of Essen University. The main purpose of MILESS is to provide students and faculties of Essen University with a digital library that supports several different functions that are needed to use multi media material for teaching and learning. Based on the IBM DB2 Digital Library product, running on an IBM RS6000/SP System, MILESS can store and retrieve multimedia digital documents in any given format; moreover, searching is possible in a very elaborate way, and access control is supported as well. MILESS is also used as the official Dissertation Online Archive of the Central Library in Essen. In December 1999 we received a Shared University Research Grant from IBM (hardware, software and services about 400.000 DM). This is the basis to extend our MILESS-services. The ongoing work is to create an "Online Digital Video Library" for education and research together with IBM. We will give an overview about the actual status of the project, some implementation details, e.g. of our Dublin Core based complex Data Model. Furthermore we will give an online demonstration of the service and we will outline some ideas about our future plans.

WWW-address: http://miless.uni-essen.de

M. Okamoto

Recognition of mathematical symbols

Formula recognition is of interest in digitization of scientific documents. We have proposed two types of methods for formula recognition. This paper describes some extensions of our methods for formulas which contain touching characters. A number of segmentation methods of touching characters in text lines have been proposed in the literatures. These methods can be applied for characters printed in only one direction; horizontal or vertical (e.g., in Japanese documents). In mathematical formulas, characters or symbols are placed in horizontal, vertical or diagonal directions. This fact means we need a new segmentation method for the touching character problem in formulas.

It is common to use some features based on vertical projection profiles for character segmentation of text lines. As a simple extension of those methods, our method use the vertical and horizontal projection profiles to segment touching characters in respective directions. However most characters or symbols in formulas are printed in slanted form, there are many cases where touching characters can be separated adequately only by a slant line not a vertical line. This line can be estimated from minimal points of the blurred image obtained by applying the Gaussian kernel to the given binary image. These points are located at the valley between blurred blob components which correspond to each character or symbol.

Since the decision whether a given blob is touching characters or not is very difficult, we use a simple criterion that if a classifier rejects the blob, we consider the blob consists of two characters. Then both of horizontal and vertical segmentation are tried, and the classification for each separated component is carried out again. Finally, the average scores for both the components are compared between two segmentations, and the one which have the higher score is adopted.

Another problem of OCR in formula recognition is that there are many symbols which have vertically or horizontally separated components. This does not cause serious problem in text recognition, because text lines are printed in only one direction. In our OCR, each separated component is classified independently at the first step, then these components are unified by using a component connectivity matrix which describes conditions in which each component is located in a symbol.

G. Schneider

Aspects of long term preservation of digital libraries

The traditional distribution of information via paper has a major advantage: access to information requires little or no technical assistance and the carrier medium for the information requires virtually no care (in a typical environment). Digital libraries provide the user with much better accessibility of information, such as instant retrieval. However, type and handling of the carrier media is changing rapidly and it is not uncommon that old storage devices can no longer be used or read.

A number of proposals have been put forward to ensure a working solution that information can be carried from one storage device to another, e.g. by robot-assisted copying. Large quantities can be thus converted at little cost.

Yet this approach does not provide a real long term solution. Experience shows that new computer systems will emerge where new software will no longer be able to read old data, even when converted. The relevant storage format has simply been forgotten. And it will be too costly to fund projects for redeveloping software reading old data formats.

Rather it is necessary to head for formats that allow the exchange of documents independently of the software. For example the DFG computer commission exchanges documents in RTF-format and the University of Gttingen is contemplating a similar suggestion for internal communication.

Before adding documents into a digital library the data format has to be standardized in a satisfactory way. It would appear that XML is a document standard with some long range validity. Probably only text-based data formats with embedded control statements will be viable. This is perhaps the most important reason why TeX has such an acceptance in the scientific community.

There are further issues to long term preservation, the functionality of software. Typically software relies on the existence of specific hardware. This is also a well-known phenomenon in other areas, like record-players being replaced by CD-players, making records more or less useless. While libraries still collect software and computer games as part of our culture it may well be that nobody will be able to play these games in the very near future. Similar problems hold for computer software in general, like data bases or teaching software, especially multimedia tools.

One option is to keep old hardware running. This will inevitably turn libraries into museums with very high operational costs. A perhaps more promising way can be emulation. Most older computer hardware, like the C64, the Gameboy, or Atari computers, can successfully and satisfactorily be emulated in software on a modern machine. New software like VMware allow the emulation of a complete operating system environment. The required information is stored in one single file, which can easily be migrated to new storage devices. The loss of performance due to software emulation can be compensated through faster and better hardware.

A feasability study has been proposed by the author to the DFG.

M. Suzuki

A Design of Optical Recognition System of Mathematical Documents An Exerimental Implimentation

In this lecture, I will describe the outline of the new experimental system of our research group[1] for the optical recognition of scientific documents including mathematical expressions. We are developping the system in collaboration with the laboratory of Toshiba Co., Ltd., producer of an OCR system "ExpressReader Pro", highly efficient in its speed and accuracy of the recognition of characters.

The system recognizes scanned images of clearly printed mathematical documents and can output the recognition results in various formats : LaTeX, MathML, etc., (and in Braille codes, since the original motivation of our work was to help visually impaired students to study and work in scientific fields).

The system is bilingual: it recognizes both Japanese and English texts. As for the mathematical expressions, the system recognizes almost all expressions used at the first courses of university mathematics level using fractions, roots, integrals, summations, limits, etc., except for matrices at present. The users are recommended to use 400DPI image scanner.

Our optical recognition function of mathematical documents is incorporated in our editor of mathematical documents named "INFTY" [2], having handwriting interface. The system looks like a simple editor or a word processor. Opening LaTeX file or MathML file, users can read and edit the texts and mathematical expressions presented on the display in a usual printed form, and can save them into any one of LaTeX, MathML and Braille format file. Thus, users can easily effectuate the transformation of data formats, for example, between LaTeX and MathML.

The displayed texts and mathematical expressions can be edited freely by using ordinary editing operations: cut, copy, paste and delete, and key inputs including LaTeX commands for mathematical expressions.

An distinctive feature of INFTY is that it provides a handwriting interface to input mathematical expressions. By clicking a button, the dialog to input pen strokes using a mouse, a data tablet or a pen display opens. The mathematical expression written by hand in this dialog box is recognized and put into the mathematical text displayed on the main board of INFTY, at its cursor position. By this handwriting interface, the system realizes a very easy intuitive methods to edit or correct mathematical expressions, requiring no special skill about, for example, LaTeX notations of mathematical expressions.

When an user opens a bitmap file or selects the scanner button in INFTY, the optical recognition function is called automatically. The recognition result of each page is displayed on the screen of computer and users can correct the errors very easily as it is mentioned above.

From each page image obtained by a scanner, the system first cuts out the text parts including mathematical expressions, table areas and picture areas. As for the pictures, the system outputs only brief informations concerning their positions and their sizes. The contents of the tables are output in separate files in CSV format.

---

[1] The members are actually R.Fukuda in Oita University, F.Tamari in Fukuoka University of Education, myself and some of their students.

[2] Infinite Type Format Document Processor in Mathematics!

F. Tamari (with S. Nouzumi, R. Fukuda and M. Suzuki)

On the segmentation of Japanese area/Mathematical area

We present a OCR engine which is able to recognize mathematical symbols and some of a part of Japanese Character(Kanji, Hiragana, Katakana) and we show that by using this engine Japanese scientific documents including mathematical symbols have been segmented into Japanese character area and mathematical expression area. Our OCR engine posses a function to reject a character except mathematical symbols including alphabets, numerals, Greek letters and other mathematical symbols used in mathematical expressions.

In our engine, 5 x 5 mesh feature and the first and the second peripheral feature of the left-hand side and the right-hand side are used as feature measure, respectively.

The dimension of a peripheral feature vector is 20. So the dimension of a feature vector in our engine is 45.

We trained about 350 characters which are mathematical symbols including alphabets, numerals, Greek letters and other mathematical notations in our engine. For each character, we calculated the feature measure of each coordinate of the feature vector. We denote a minimum and a maximum value of the measure by a_i and b_i, respectively for $i = 1, 2, \ldots, 45$. We set interval $I = [a_i - \alpha_i, b_i + \beta_i]$, where α_i and β_i are constant numbers to recognize a input character that have been not yet trained.

When circumscribed rectangular image of a character is scanned from the left-hand side(top) to the right-hand sid(bottom), we count the number of the places that change from a white pixel to a black pixel. The Horizontal Cross Count (Vertical Cross Count) is a maximum value of such numbers and is denoted by HCC(VCC).

Next, for each coordinate of a feature vector, i.e., the first coordinate, the second coordinate, ... , the 45 th coordinate, we check if the feature measure of an inputed pattern is in the dictinary. If for some coordinate, the feature measure of the pattern is not in the dictionary, the pattern is thought as a data except characters subject to the recognition in our engine and it is say that the character is rejected.

We describe how to segment Japanese scientific documents including mathematical expresstions into Japanese character area and mathematical expression area. It is noticed that there are some geometric differences between Japanese characters and mathematical symbols. We regist the following geometric features of a character in a geometirc dictionary:

(1) AspectArg (2) HCC (3) VCC (4) Maximum value of connected components of a character. (5) Space in a Japanese character (6) Left-hand side of a Japanese character (7) Right-hand side of a Japanese character.

AspectArg is arctangent(height of the rectangular frame enclosing a character/width of the rectangular frame enclosing a character).

A mathematical expression line is one or a few of ordinary lines. So an ordinary line and a line that represents mathematical expressions is called a sentence line. If a sentence line is divided into several parts by spaces, i.e., projections onto the X axis, we get a block consists of a part of Japanese character or a Japanese character or a mathematical symbol or a few of mathematical symbols, i.e., summation, integration and so on. That block is called a simple rectangle. We consider a graph for their bloks and spaces among characters in a sentence line. We take a space as

a node and between a node and another node is connected by two arcs. One is used for Japanese OCR engine and it is called a Japanese arc and the other is used for mathematical OCR engine and it is called a mathematical arc. We give a cost to a Japanese and a mathematical arc, respectively. A path in a sentence line consists of several Japanese arcs and mathematical arcs. We consider all paths in the line and search a path that minimizes the total cost in the path , where a total cost is summed by costs for Japanese arcs and mathematical arcs.

We used DP(Dynamic Programming) algorithm to solve the problem of searching a path which have a minimum cost.

We give costs for a mathematical arc as follows:

: (1) First we initialize a cost, i.e., we set cost = 100.
: (2) If the result of recogniton of a inputed data in the Japanese engine is equal to that in the mathematical engine, then we subtract 30 from the cost.
: (3) If the result of recognition of a inputed data is rejected, we check the number of connected components which are rejected and not rejected.
  : (a) We add 2 times the number of connected components whici are rejected in the simple rectangle to the cost.
  : (b) We subtract the number of connected components which are not rejected in the rectangle to the cost.
  : (c) If the number of connected components which are rejeced is greater than or equal to one half of the number of connected components in the simple rectangle, we add 35 to the cost.
  : (d) If a rejected simple rectangle consists of a connected component , we add 30 to the cost.
: (4) If the result of recognition of a inputed data is not rejected, we compare geometric features of the data,i.e., HCC, VCC and so on with those of its recognition result that is registered in the geometric dictionary.
  For exsample,
  : (a) Let us denote the difference between AspectArg of a inputed pattern and that of its recognition result in the dictionary by d. We add the absolutely of d to the cost.
  : (b) If HCC(VCC) of a inputed data is greater than that of the dictionary, then we add 4 times the absolutely of the difference of them to the cost, else we add 2 times the absolutely of the difference of them to the cost.

We omit how to give costs for a Japaneses arc.

Furthermore we used some grammatic methods about Japanese in order to segment the documents.

S.W. Thomas

JSTOR's experience with retrodigitization of mathematics literature

Introduction

JSTOR (http://www.jstor.org) is a multi-disciplinary electronic archive of scholarly journals. Each journal in the archive has been digitized in its entirety, starting with the first issue and continuing almost to the present. All titles in JSTOR have been selected as being significant in their disciplines. The archive is available to users at participating institutions through the WWW.

Each title in JSTOR is digitized in 3 forms. (1) Each page is individually scanned at high resolution (600 DPI). (2) Each page image is converted to text by optical character recognition (OCR). (3) Metadata for the journal and its contents, such as article title, authors, abstract, etc. is hand-typed to ensure accuracy.

JSTOR presents page images to the user in several formats, at low resolution for viewing on-screen and at high resolution for printing. The OCR text is used to provide a full-text search capability. It is sufficiently accurate for searching (99.95

The multi-disciplinary nature of JSTOR appears to be of value. In a recent week, 70% of searches involved more than one discipline, and almost 60% more than two.

Why mathematics?

JSTOR began digitizing mathematical journals early in its history (1996). Mathematics was the fourth discipline added to the JSTOR collection. The JSTOR collection currently includes 11 Mathematics titles and another11 in Statistics.

Mathematics has a long history of scholarly publication, and the older literature is often still relevant and consulted. The Science and Mathematics librarian at the University of Michigan informed us that the library keeps the full run of many Mathematics journals on the shelves in the open stacks. The older issues of journals in most other disciplines tend to be in closed stacks. These qualities made Mathematics a natural choice for JSTOR, with our emphasis on providing access to the older material.

Usage has borne out this decision. The average "age" of the articles that are read by users in most JSTOR disciplines is 3-12 years, but in Mathematics it is over 30 years!

Challenges and Solutions

However, digitizing Mathematics literature presents challenges beyond those encountered in digitizing the Humanities titles, such as History. In particular, the basically graphical nature of mathematical formulas presents a significant hurdle to text-based digital forms. JSTOR's emphasis on page images meant that we would have no problem in faithful visual reproduction of content. But it was not immediately clear what to do with the metadata and OCR text. Our decisions on how to textually represent mathematical notation in the metadata and OCR text were guided by the uses to which that content is put in JSTOR and by JSTOR's overall mission. We felt a tension between providing "as good as possible" digitization, and our requirement that the full multi-disciplinary collection fit together cohesively.

The OCR text is used for full-text searching. Since search terms are (currently) entered as text, it is not possible to enter a mathematical formula, thus we do not

need to capture the form of the formula in the OCR text. (But see Open Questions, below.)

The metadata, on the other hand, is presented on-screen in citations, tables of contents, and lists of search results. We must therefore capture, as well as possible, the mathematical content of titles, abstracts, etc. After looking at other online mathematics systems, and consulting with local specialists, we elected to capture mathematical content using the AMS LaTeX encoding. LaTeX encodes mathematics in a "plain text" form, and is thus compatible with the "plain text" used in other disciplines, with the data entry systems used by our suppliers, and with the full-text search subsystem.

At the same time, LaTeX is comprehensible to the target audience of mathematicians, as many mathematics journals require that papers be written for submission using LaTeX. Because of this, we felt that presenting the "raw" LaTeX in citations was appropriate for the field.

Recent improvements

As with any tool, once you have it, you start using it for purposes other than that for which it was originally intended. Thus it was with LaTeX in JSTOR. Soon, LaTeX coding was appearing in the metadata for journals in non-mathematical fields such as Economics and General Science. But readers in these disciplines were not necessarily familiar with LaTeX, and would query the "bizarre-looking" text. We entered into an investigation of ways to render LaTeX coded material into HTML for on-screen display. Our goals included the following: to eliminate LaTeX markup for non-mathematical uses, such as symbol inclusion, superscripts and subscripts; to faithfully render as much mathematical notation as possible into HTML; and to "fail gracefully" when presented with complex LaTeX coding. Other desirable features were that the HTML be self-contained, and that it be as accessible as possible to people with visual impairments.

One method that many sites use for rendering mathematics in HTML is to place small images inline to represent of symbols and bits of mathematical notation. The advantage of this method is that it is possible to display arbitrary notation. The drawbacks include that the rendered images are of a particular size, font, and color, which may not match the user's choices; and that when text is copied from the page, there is no text corresponding to the images. The former is a particular problem for accessibility to those with visual impairments.

Web browser "plug-ins" are available that will render TeX. We did not want to require casual users to install a plug-in before being able to use JSTOR, and therefore determined not to use this solution.

We settled on a hybrid solution, using Unicode character encoding for special symbols, and preprocessing the LaTeX encoded formulas into HTML representation in the metadata. The preprocessing proceeds in the following steps:

1. If the item (title, abstract, etc.) contains no LaTeX coding, stop.
2. Perform simple character entity translations to the corresponding Unicode character entities. Thus, for example, "`$\alpha$`" is translated to "α". If at the end of this process, no LaTeX coding remains, stop.
3. Use the program "TtH" to translate each remaining LaTeX expression to HTML. If TtH fails to translate an expression, leave that expression as LaTeX in the rendered output.
4. If any transformations were done in steps 2 or 3, save the output.

We began deploying this solution in mid-June, 2000.

Open questions

We are confident that the capture of mathematical content in the JSTOR metadata is reasonably complete. However, there remain many open and interesting questions and areas for potential improvement. One question is how to capture mathematical content in the full text created by OCR. A related question is how to specify and execute searches for mathematical formulas, especially given that two scholars might encode the same equation in two textually different, but equally valid ways.

As we consider possible extensions of the ways in which we capture, use, and display mathematics, we at JSTOR must also ensure that we do not lose the cross-disciplinary nature of the JSTOR archive. We will continue to work with the sometimes conflicting goals of providing special service for a particular discipline, while continuing to provide uniformity of service for all.

K. Yokota

Layout analysis of mathematical documents

Toshiba is producing some types of OCR and scanners. First we introduce such products of OCR. And then, we shows some our technologies for OCR. Next, we explain the joint research with professor Suzuki's group to achieve the recognition of mathematical documents. Finally we suggests a new method to detect mathematical formulas based on a basic natural language processing.

From text/mathematics area, the system separates the true text parts and the mathematical expression parts using the recognition results of characters together with their positions and their sizes. In the case of English documents, linguistic informations such as English words dictionary, mathematical technical terms dictionary and some morphological analysis are used, in order to improve the accuracy of the recognition of words and the separation of text parts and mathematical expression parts.

As for the recognition of mathematical expressions, we introduced a new algorithm to improve the robustness of the recognition against the variation of the styles of printing of the documents, using the spanning tree of minimum cost in the network generated by virtual links of characters. The details of this algorithm and experimental results are described in Y. Etou's lecture.

Knowledge - information
what we know ourselves
What we can find out
What we have a copy of...
Copy available to link away
URL

digital library

zblmath.fiz-karlsruhe.de
jfm@zblmath.fiz-karlsruhe.de
Jahrbuch project
Jahrbuch über die Fortschritte der Mathematik

JSTOR - good stuff -
present (limited) math
in new HTML <sml>
etc - <smp>

North Rhein/Westfalia (NRW)
Bielefeld.
www.digibib-nrw.de

Dieper: international register of digitized periodicals

www.doi.org
www.crossref.org
document object identifier
(go to SIGSAM?)

Suzuki @ math.kyushu-u.ac.jp
uses doshisha language, handwriting of papers/edited + custom built math characters + plane

Identifying Japanese by maximum horizontal/vertical crossings (eliminating high counts for noise e.g. ... ← rare high count) b/w ratio, other measures.

RK Dennis
Mathnet
1.6 million
2-5 million pages
Since printing
50 million pages
Meta data - Reviews
Citation
Users Comments
Author info
Links
People

(example (NASA funded)
astrophysics, with path permanent
ads www.harvard.edu

Search features
patterns
concepts
symbols
Results - translate, print, compute
verify

American Institute of Math
www.aimath.org
John Fry of Fry's Electronics

project Euclid at Cornell
Mellon Fdn - provide
online support for small journals

Math hubs Cornell - Göttingen - Michigan

Where to go?
Coordination Web Site -
partnerships, funding, list
of work already done.

Diver library funds
to doing OCR of journals
at $2/page. Get 100
universities → .02/page

example of simplified notations in
group theory.